Yash Akhauri

GitHub | LinkedIn | akhauri.yash@gmail.com | +91 78915 12802

A Research Scientist specializing in Neural Network optimization with a keen interest in Mathematics and Physics.

EDUCATION

BITS PILANI | B.E. in Electronics and Instrumentation
Aug 2016 - May 2020 | RJ, India
Fluent: Python3, PyTorch, Mathematica
Familiar: C++, Java, CUDA, OpenMP, Tensorflow, Android Studio, LibGDX, Docker

EXPERIENCE

INTEL | Research Scientist — May 2020 – Present, Bangalore, India

- Research Scientist at the Cloud Systems Research (CSR) Lab in the Systems and Software Research (SSR) Group.

XILINX RESEARCH | Visiting Scholar — Aug 2019 – May 2020, Dublin, Ireland

- Developed a library for co-design of neural network topologies and reconfigurable hardware that maps to an efficient FPGA implementation without the need for a custom accelerator architecture or a scheduler.
- Targeted the Jet Substructure Classification task as part of CERN LMS L1 trigger experiments, used the library to deploy models with $10\times$ lower latency than FPGA4HEP designs.Demonstrated quantization library Brevitas to CERN `[GitHub]`.

URANIOM | Research Intern — Jan 2019 – Jul 2019, France (Remote)

- Implementing semantic segmentation models for face transfer across GIFs, progressive GANs for realistic UV map generation and explored effective weight-sharing strategies for neural networks under a research collaboration.

WOLFRAM | Undergraduate Researcher — June 2018 – July 2018, Massachusetts

- Developed HadaNet MLPs in the Wolfram Language and worked on C OpenMP kernels for GEMM and Convolutions using the Hadamard Binarization algorithm. `[OpenMP kernel]` | `[CUDA kernel]` | `[Whitepaper]`

RESEARCH

PAPERS

LogicNets: Co-Designed Neural Networks and Circuits for Extreme-Throughput Applications
`[IEEE [1]]` Yash Akhauri*, Yaman Umuroglu*, Nicholas J. Fraser, Michaela Blott — *FPL'20* | Sweden | Sept 2020
Paper accepted at FCCM'20 as a poster presentation. Video presentation can be found here.

High-Throughput DNN Inference with LogicNets
`[IEEE [1]]` Yaman Umuroglu, Yash Akhauri, Nicholas J. Fraser, Michaela Blott — *FCCM'20* | AR, USA | May 2020

HadaNets: Flexible Quantization Strategies for Neural Networks
`[IEEE]` Yash Akhauri — *CVPR'19 Workshop* | CA, USA | Jun 2019
Paper accepted at CVPR'19 UAVision workshop - Orals .
Delivered a "Theatre Talk" and poster at the Intel Demo Booth at CVPR'19.

Exposing Hardware Building Blocks to Machine Learning Frameworks
`[ArXiv -- Bachelor's Thesis]` Yash Akhauri — Dec 2019

[1] To be published
[1] To be published

TALKS

Wolfram Technology Conference

Speaker — Champaign, IL | Oct. 2018

Delivered a talk on my research on Hadamard Neural Networks.

Intel AI Meetup

Speaker — Delhi, IN | Sept. 2018

[PPTX] [Article] Spoke about my research on scaling AI using Intel technologies. This event was organized by Intel.

Intel AI DevCon

Poster — San Francisco, Bangalore | May & Aug 2018

Presented posters on quantized GEMM kernels for Intel Xeon Phi

GRANTS

Intel Nervana Early Innovators Grant — $5000

Received research grant to develop Binary Precision Neural Networks and Real time Artistic Style Transfer. The technical article can be found [here.] The code can be found [here.]
Intel AI Academy Success Story [Link] published by Intel for the research done as part of this grant in the field of Quantized Neural Networks.

Intel CVPR Travel Grant — $3000

Received a travel grant from Intel to present research at the Intel Demo Booth at CVPR'19.

Wolfram Student Aid — $2400

Received aid to attend the Wolfram Summer School and develop Hadamard Binary Neural Networks.

KVPY Scholar

Selected as a KVPY scholar by the Department of Science and Technology, Government of India.

INSPIRE Scholarship

Selected for the INSPIRE Scholarship by the Department of Science and Technology (DST), Government of India.

PROJECTS

Exploiting Huffman Coding and Weight-Sharing for Memory-Efficient Inference on FPGAs. PyTorch

Developed a neural network weight sharing strategy and proposed a methodology to leverage this strategy in the FINN architecture for Multi-Layer Offload and Data-Flow Architectures of Neural Network deployment on FPGAs.

Whitepaper - Improving distributed mesh computing with Hadamard Binary Neural Networks. [Link]

xGEMM & xCONV — C++, CUDA, OpenMP

Coded efficient 3D convolutional and GEMM kernels for XNOR (bit quantized) networks using CUDA C programming and OpenMP. Optimized kernels are for Intel processors and Nvidia GPUs. Invited to present a poster at Intel AI DevCon and Intel AI Student Ambassador Summit, San Francisco. The codes can be found here: [OpenMP kernel] | [CUDA kernel].

Real time artistic style transfer — Python, Tensorflow, OpenCV

/

GravDash — Java, LibGDX, Android Studio

Developed an android game using Java and libGDX as the framework.

Blog — Python, Java, C++, Tensorflow, OpenMP

Maintaining a [blog] covering various AI related topics with over 10000 hits.